Domain Adaptation of Majority Votes via Perturbed Variation-based Label Transfer

Emilie Morvant*
Institute of Science and Technology Austria
Klosterneuburg, 3400 Austria
emorvant@ist.ac.at

Abstract

We tackle the PAC-Bayesian Domain Adaptation (DA) problem [1]. This arrives when one desires to learn, from a source distribution, a good weighted majority vote (over a set of classifiers) on a different target distribution. In this context, the disagreement between classifiers is known crucial to control. In non-DA supervised setting, a theoretical bound – the C-bound [2] – involves this disagreement and leads to a majority vote learning algorithm: MinCq [3]. In this work, we extend MinCq to DA by taking advantage of an elegant divergence between distribution called the Perturbed Varation (PV) [4]. Firstly, justified by a new formulation of the C-bound, we provide to MinCq a target sample labeled thanks to a PV-based self-labeling focused on regions where the source and target marginal distributions are closer. Secondly, we propose an original process for tuning the hyperparameters. Our framework shows very promising results on a toy problem.

1 Introduction

Nowadays, due to the expansion of Internet a large amount of data is available. Then, an important issue in Machine Learning is to develop methods able to transfer knowledge from different information sources or tasks, which is known as Transfer Learning (see [5] for a survey). In this work, we tackle the hard [6] problem of unsupervised Domain Adaptation (DA), which arises when we want to learn from a distribution – the source domain – a well performing model on a different distribution – the target domain – for which one has an unlabeled sample. Consider, for instance, the common problem of spam filtering, in which one task consists in adapting a model from one user to a new one. One popular solution is to take advantage of a divergence between the domains, with the intuition that we want to minimize the divergence while preserving good performance on the source data [7, 8, 1]. Some classical divergences involve the disagreement between classifiers, which appears crucial to control. Another divergence, the Perturbed Variation (PV) [4], is based on this principle: Two samples are similar if every target instance is close to a source instance. In this work, we focus on the PAC-Bayesian DA setting introduced in [1] for learning a good target weighted majority vote over a set of classifiers (or voters). A key point is that the divergence used, which takes into account the expectation of the disagreement between pairs of voters, is justified by a recent tight bound on the risk of the majority vote: the C-bound [2]. This C-bound leads to an elegant and well performing algorithm for supervised classification, called MinCq [3]. Our contribution consists in extending MinCq to the DA scenario, thanks to a label transfer from the source domain to the target one. First, we propose in section 3 a new version of the C-bound suitable for every label transfer defined by a label function. Then, we design in section 4 such a function thanks to the empirical PV. Concretely, our PV-based label transfer focuses on the regions where the source and target marginals are closer, and labels the (unlabeled) target sample only in these regions. Afterwards, we provide to MinCq

*This work was in parts funded by the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

this auto-labeled target sample. We also make use of the PV to define an original hyperparameters validation. Finally, we show empirically in section 5 that our approach implies good and promising results on a toy problem, better than a nearest neighboorhood-based transfer.

2 Notations and Background

Throughout this paper, we consider the PAC-Bayesian DA setting described in [1] for classification tasks where $X \in \mathbb{R}^d$ is the input space of dimension d and $Y = \{-1, +1\}$ is the label set. The source domain P_S and the target domain P_T are two different distributions over $X \times Y$. D_S and D_T are the respective marginal distributions over X. In the PAC-Bayesian theory, introduced in [9], given a set of classifiers (that we called voters) $\mathcal{H}$ from X to $\mathbb{R}$ and given a prior distribution π of support $\mathcal{H}$, the learner aims at finding a posterior distribution ρ leading to a ρ-weighted majority vote B_ρ over $\mathcal{H}$ with good generalization guarantees. B_ρ is defined as follows.

Definition 1. *Let $\mathcal{H}$ be a set of voters from X to $\mathbb{R}$. Let ρ be a distribution over $\mathcal{H}$. The ρ-weighted majority vote B_ρ (sometimes called the Bayes classifier) is,*

$$\forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

The true risk of B_ρ on a domain P is, $\mathbf{R}_P(B_\rho) = \frac{1}{2}\big(1 - \mathbf{E}_{(\mathbf{x},y)\sim P}\, yB_\rho(\mathbf{x})\big)$.

Usual PAC-Bayesian generalization guarantees (*e.g.* [10, 11, 12, 13, 14]) bound the risk of the stochastic Gibbs classifier G_ρ, which labels an example $\mathbf{x}$ by first drawing a voter h from $\mathcal{H}$ according to ρ, then returns $\operatorname{sign}[h(\mathbf{x})]$. The risk of G_ρ corresponds to the expectation of the risks:

$$\mathbf{R}_P(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_P(h) = \tfrac{1}{2}\big(1 - \mathop{\mathbf{E}}_{h\sim\rho}\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} yh(\mathbf{x})\big).$$

It is then easy to relate B_ρ and G_ρ by: $\mathbf{R}_P(B_\rho) \leq 2\mathbf{R}_P(G_\rho)$.
In that light, the authors of [1] have done a PAC-Bayesian analysis of DA. Their main result is stated in the following theorem.

Theorem 1 ([1])**.** *Let $\mathcal{H}$ be a set of voters. For every distribution ρ over $\mathcal{H}$, we have,*

$$\mathbf{R}_{P_T}(G_\rho) \ \leq\ \mathbf{R}_{P_S}(G_\rho) \ +\ \operatorname{dis}_\rho(D_S, D_T) \ +\ \lambda_\rho,$$

where λ_ρ is a term related to the true labeling on the two domains[1]*, and*
$\operatorname{dis}_\rho(D_S, D_T) = \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left(\mathop{\mathbf{E}}_{\mathbf{x}_t\sim D_T} h(\mathbf{x}_t)h'(\mathbf{x}_t) - \mathop{\mathbf{E}}_{\mathbf{x}_s\sim D_S} h(\mathbf{x}_s)h'(\mathbf{x}_s)\right)\right|$ *is the domain disagreement.*

This bound reflects the philosophy in DA: It is well known [7] that a good adaptation may be possible if the divergence between the domains is small while achieving good performance on the source domain. The point which calls our attention in this result is the definition of the domain disagreement, $\operatorname{dis}_\rho(D_S, D_T)$, directly related to the disagreement between pairs of voters, and justified by the definition of the following theoretical bound called the C-bound [3, 2].

Theorem 2 (The C-bound as expressed in [3])**.** *For all distribution ρ over $\mathcal{H}$, for all domain P_S over $X \times Y$ of marginal (over X) D_S, if $\mathop{\mathbf{E}}_{h\sim\rho}\ \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P_S} y_s h(\mathbf{x}_s) > 0$, then,*

$$\mathbf{R}_{P_S}(B_\rho) \leq 1 - \frac{\left(\mathop{\mathbf{E}}_{h\sim\rho}\ \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P_S} y_s h(\mathbf{x}_s)\right)^2}{\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\ \mathop{\mathbf{E}}_{\mathbf{x}_s\sim D_S} h(\mathbf{x}_s)h'(\mathbf{x}_s)}.$$

Since we can remark the C-bound's denominator is also related to the disagreement between pairs of voters, we propose, in the next section, a new formulation suited for DA. Before, we recall the supervised classification algorithm MinCq [3] which ensues from the C-bound (and described in Algo. 1). Concretely, MinCq learns a performing majority vote by optimizing the empirical counterpart of the C-bound: It minimizes the denominator, *i.e.* the disagreement (Eq. (1)), given a fixed numerator *i.e.* a fixed margin for the majority vote (Eq. (2)), under a particular regularization (Eq. (3)).[2] Note that its consistency is justified by a PAC-Bayesian generalization bound.
Since the C-bound, and thus MinCq, focus on the disagreement between voters, which is crucial to control in DA [7, 8, 1], we propose to make use of the C-bound and MinCq in a DA perspective.

[1] Since one usually omits this term in algorithms, we do not develop it. More details could be found in [1].
[2] For more technical details on MinCq, please refers to [3].

Algorithm 1 MinCq($S, \mathcal{H}, \mu$)

input A sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|S|}$, a set of voters $\mathcal{H}$, a desired margin $\mu > 0$

output $B_\rho(\cdot) = \text{sign}\left[\sum_{j=1}^{|\mathcal{H}|}\left(2\rho_j - \frac{1}{|\mathcal{H}|}\right)h_j(\cdot)\right]$

$$\textbf{Solve} \quad \underset{\boldsymbol{\rho}}{\text{argmin}} \ \boldsymbol{\rho}^T \mathbf{M}\boldsymbol{\rho} - \mathbf{A}^T\boldsymbol{\rho}, \tag{1}$$

$$\textbf{s.t.} \ \mathbf{m}^T\boldsymbol{\rho} = \frac{\mu}{2} + \frac{1}{2|S||\mathcal{H}|}\sum_{j=1}^{|\mathcal{H}|}\sum_{i=1}^{|S|} y_i h_j(\mathbf{x}_i), \tag{2}$$

$$\forall j \in \{1, \ldots, |\mathcal{H}|\}, \quad 0 \leq \boldsymbol{\rho}_j \leq \frac{1}{|\mathcal{H}|}, \tag{3}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{|\mathcal{H}|})^T$ is a vector of weights,

$\mathbf{M}$ is the $|\mathcal{H}|\times|\mathcal{H}|$ matrix formed by $\sum_{i=1}^{|S|}\frac{h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)}{|S|}$ for $(j, j') \in \{1, \ldots, |\mathcal{H}|\}^2$, and:

$$\mathbf{m} = \left(\tfrac{1}{|S|}\sum_{i=1}^{|S|} y_i h_1(\mathbf{x}_i), \ldots, \tfrac{1}{|S|}\sum_{i=1}^{|S|} y_i h_{|\mathcal{H}|}(\mathbf{x}_i)\right)^T$$

$$\mathbf{A} = \left(\sum_{j=1}^{|\mathcal{H}|}\sum_{i=1}^{|S|}\frac{h_1(\mathbf{x}_i)h_j(\mathbf{x}_i)}{|\mathcal{H}||S|}, \ldots, \sum_{j=1}^{|\mathcal{H}|}\sum_{i=1}^{|S|}\frac{h_{|\mathcal{H}|}(\mathbf{x}_i)h_j(\mathbf{x}_i)}{|\mathcal{H}||S|}\right)^T$$

Algorithm 2 $\widehat{PV}(S, T, \epsilon, d)$

input $S = \{\mathbf{x}_s\}_{s=1}^{|S|}$ and $T = \{\mathbf{x}_t\}_{t=1}^{|T|}$ are unlabeled samples, $\epsilon > 0$, a distance d

output $\widehat{PV}(S, T)$

1. $G \leftarrow \big(V = (A, B), E\big)$, where $A = \{\mathbf{x}_s \in S\}$ and $B = \{\mathbf{x}_t \in T\}$, $e_{st} \in E$ if $d(\mathbf{x}_s, \mathbf{x}_t) \leq \epsilon$
2. $M_{ST} \leftarrow$ Maximum matching on G
3. $S_u \leftarrow$ number of unmatched vertices in S
   $T_u \leftarrow$ number of unmatched vertices in T
4. Return $\widehat{PV}(S, T) = \frac{1}{2}\left(\frac{S_u}{|S|} + \frac{T_u}{|T|}\right)$

Algorithm 3 PV-MinCq($S, T, \mathcal{H}, \mu, \epsilon, d$)

input $S = \{(\mathbf{x}_s, y_s)\}_{s=1}^{|S|}$ a source sample, $T = \{\mathbf{x}_t\}_{t=1}^{|T|}$ a target sample, $\mathcal{H}$, $\mu > 0$, $\epsilon > 0$, d

output $B_\rho(\cdot)$

$M_{ST} \leftarrow$ Step 1. and 2. $\widehat{PV}(S, T, \epsilon, d)$

$\widehat{T} \leftarrow \{(\mathbf{x}_t, y_s) : (\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}, (\mathbf{x}_s, y_s) \in S\}$

return MinCq($\widehat{T}, \mathcal{H}, \mu$)

3 A C-bound suitable to Domain Adaptation with Label Transfer

First, we propose to rewrite the C-bound with a labeling function $l : X \mapsto Y$, which associates a label $y \in Y$ to an unlabeled example $\mathbf{x}_t \sim D_T$. Given such a function, the C-bound becomes:

Corollary 3. *For all distribution ρ over $\mathcal{H}$, for all domain P_T over $X \times Y$ of marginal (over X) D_T, for all labeling functions $l : X \mapsto Y$ such that $\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{\mathbf{x}_t\sim D_T} l(\mathbf{x}_t)h(\mathbf{x}_t) > 0$, we have,*

$$\mathbf{R}_{P_T}(B_\rho) \leq 1 - \frac{\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{\mathbf{x}_t\sim D_T} l(\mathbf{x}_t)h(\mathbf{x}_t)\right)^2}{\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\mathop{\mathbf{E}}_{\mathbf{x}_t\sim D_T} h(\mathbf{x}_t)h'(\mathbf{x}_t)} + \frac{1}{2}\left|\mathop{\mathbf{E}}_{(\mathbf{x}_t,y_t)\sim P_T}(y_t - l(\mathbf{x}_t))\right|.$$

The first two terms correspond simply to the usual C-bound measured with the labeling function l. The term $\frac{1}{2}\left|\mathbf{E}_{(\mathbf{x}_t,y_t)\sim P_T}(y_t - l(\mathbf{x}_t))\right|$ can be seen as a divergence between the true labeling and the one provided by l: The more similar l and the true labeling are, the tigher the bound is.

With a DA point of view, an important remark is that only one domain appears in this bound. Then, we guess that this domain is the target one, and that the computation of a relevant labeling function has to make use of the information carried by the source labeled sample S. Concretely, given a labeled source instance $(\mathbf{x}_s, y_s)$, we want to transfer its label y_s to an unlabeled target point $\mathbf{x}_t$ close to $\mathbf{x}_s$. This will give rise to an auto-labeled target sample, on which we can apply MinCq. To tackle the issue of defining the label transfer, we propose, in the following, to investigate a recent measure of divergence between distributions: the Perturbed Variation [4].

4 A Domain Adaptation MinCq with the Perturbed Variation

We first recall the definition of the Perturbed Variation (PV) proposed in [4].

Definition 2 ([4])**.** *Let D_S and D_T two marginal distributions over X, let $M(D_S, D_T)$ be the set of all joint distributions over $X \times X$ with marginals D_S and D_T. The perturbed variation w.r.t. a distance $d : X \times X \mapsto \mathbb{R}$ and $\epsilon > 0$ is defined by,*

$$PV(D_S, D_T, \epsilon, d) = \inf_{\mu\in M(D_S,D_T)} \mathop{\mathbf{Pr}}_{\mu}\left[d(\mathcal{X}, \mathcal{X}') > \epsilon\right],$$

over all pairs $(D_S, D_T) \sim \mu$, such that the marginal of $\mathcal{X}$ (resp. $\mathcal{X}'$) is D_S (resp. D_T).

In other words, two samples are similar if every target instance is close to a source instance. Note that this measure is consistent and that its empirical counterpart $\widehat{PV}(S, T)$ can be efficiently computed by a maximum graph matching procedure described in Algo. 2 [4].

In our label transfer objective, we then propose to make use of the maximum graph matching computed M_{ST} by the PV at step 2 of Algo. 2 (with d the euclidian distance and ϵ a hyperparameter).

Target rotation angle	20°	30°	40°	50°	60°	70°	80°
MinCq	92.1	78.2	69.8	61	50.1	40.7	32.7
SVM	89.6	76	68.8	60	47.18	26.12	19.22
TSVM	**100**	78.9	74.6	70.9	64.72	21.28	18.92
DASVM	**100**	78.4	71.6	66.6	61.57	25.34	21.07
PBDA	90.6	89.7	77.5	58.8	42.4	37.4	39.6
DASF	98	92	83	70	54	43	38
NN-MinCq	97.7	83.7	77.7	69.2	58.1	47.9	42.1
PV-MinCq	99.9	**99.7**	**99**	**91.6**	**75.3**	**66.2**	**58.9**

Table 1: Average accuracy results on 10 runs for 7 rotation angles.

Concretely, we label the examples from the unlabeled target sample T with M_{ST}, with the intuition that if $\mathbf{x}_t \in T$ belongs to a pair $(\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}$, then $\mathbf{x}_t$ is affected by the true label of $\mathbf{x}_s$. Else, we remove $\mathbf{x}_t$ from T. The auto-labeled sample obtained is denoted by $\widehat{T}$. Then we provide $\widehat{T}$ to MinCq. Our global procedure, called PV-MinCq, is summarized in Algo. 3.
Obviously, a last question concerns the hyperparameters selection. Usually in DA, one can make use of a reverse/circular validation as done in [15, 1, 16]. However, since in our specific situation with PV-MinCq, we have not directly make use of the value of the PV, we propose to select parameters with a k-fold validation process optimizing the trade-off: $\mathbf{R}_S(B_\rho) + \widehat{PV}(S,T)$, where $\mathbf{R}_S(B_\rho)$ is the empirical risk on the source sample. This heuristic is justified by the philosophy of DA: Minimize the divergence (measured with the PV) between the domains while keeping good performances on the source labels transferred on the target points.

5 Experimental Results

We tackle the toy problem called "inter-twinning moon", each moon corresponds to one class. We consider seven target domains rotating anticlockwise the source domain according to 7 angles. Our PV-MinCq is compared with MinCq and SVM with no adaptation, and with DA approaches: The semi-supervised Transductive-SVM (TSVM) [17], the iterative DA algorithms DASVM [15] (based on an auto-labeling) and DASF [16] (based on the usual bound in DA [7]), and the PAC-Bayesian DA method PBDA [1]. We also report a version of MinCq that makes use of a k-NN based auto-labeling (NN-MinCq): We label a target point with a k-NN classifier of which the prototypes comes from the source sample. We used a Gaussian kernel for all the methods. The preliminary results – illustrated on Tab. 1 – are very promising. Firstly, PV-MinCq outperformns on average the others, and appears more robust to change of density (NN-MinCq and MinCq appears also more robust). This confirms the importance to take into account the disagreement between voters in DA[3]. Secondly, the PV-based labeling implies better results than the NN one. Unlike a NN-based labeling, using the matching implied by the computation of the PV appears to be a colloquial way to control the divergence between domains since it clearly focuses on high density region by removing the target example without matched source instance, in other words on regions where the domains are close. These two points confirm that the PV is a relevant measure to control the process for a DA task.

6 Conclusion and Future Work

In this work, we have proposed a first procedure to tackle DA by making use of the recent algorithm called MinCq. Indeed, MinCq allows us to take into account the disagreement between classifiers, which is known to be crucial in DA. Our approach has the originality to directly minimize a risk on the target domain thanks to a labeling defined with the Perturbed Variation distance between distributions. The preliminary results obtained are promising, and we would like to apply the method to real-life applications. Another exciting perspective is to define new label transfer functions, for example by computing the PV with a more adapted distance d such as the domain disagreement.

[3]Note that preliminary experiments using PV with a SVM have implied poor results. This also probably confirms the importance of the disagreement.

References

[1] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. PAC-Bayesian domain adaptation bound with specialization to linear classifiers. In *Proceedings of International Conference on Machine Learning*, 2013.

[2] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Proceedings of Annual Conference on Neural Information Processing Systems*, 2007.

[3] F. Laviolette, M. Marchand, and J.-F. Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *Proceedings of International Conference on Machine Learning*, June 2011.

[4] M. Harel and S. Mannor. The perturbed variation. In *NIPS*, pages 1943–1951, 2012.

[5] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[6] S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of Algorithmic Learning Theory*, pages 139–153, 2012.

[7] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 137–144, 2007.

[8] Yishay Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 1041–1048, 2008.

[9] D. A. McAllester. PAC-bayesian model averaging. In *Proceedings of annual conference on Computational learning theory*, pages 164–170, 1999.

[10] D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

[11] M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

[12] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

[13] O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Institute of Mathematical Statistic, 2007.

[14] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of International Conference on Machine Learning*, 2009.

[15] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[16] E. Morvant, A. Habrard, and S. Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, 2012.

[17] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of International Conference on Machine Learning*, pages 200–209, 1999.